Learning with Ambiguities

Amir Saffari

Oxford Brookes University
Vision R&D, Sony Computer Entertainment Europe, London

8 March, 2011
Outline

1. Semi-Supervised Learning
2. Self-Training
3. Generative Models
4. Margin Assumption
5. Cluster and Manifold Assumption
6. Multi-View Learning
7. Related Topics
Supervised learning is all about finding mappings from input (feature) space to output space:

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

\[ (x_2, -1) \]

\[ (x_1, +1) \]
Supervised learning is all about finding mappings from input (feature)
space to output space:

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]

\[(x_1, +1)\]

\[(x_2, -1)\]
Supervised learning is all about finding mappings from input (feature) space to output space:

\[ f(x; \theta) \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y} \]

\[(x_1, +1)

\[(x_2, -1)\]
In **Semi-supervised learning** we wish to find mappings by using both labeled and unlabeled data:

\[ f(x; \theta) \in \mathcal{F} : \mathcal{X} \to \mathcal{Y} \]
Why Should SSL Make Sense?

Learner

\[ f(x; \theta) \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y} \]

Labeled data

\[ D_l = \{ (x, y) \in \mathcal{X} \times \mathcal{Y} \} \]
Why Should SSL Make Sense?

Learner

\[ f(x; \theta) \in F : \mathcal{X} \rightarrow \mathcal{Y} \]

Labeled data

\[ D_l = \{ (x, y) \in \mathcal{X} \times \mathcal{Y} \} \]

Unlabeled data

\[ D_u = \{ x \in \mathcal{X} \} \]
Why Should SSL Make Sense?

Learner

\[ f(x; \theta) \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y} \]

Labeled data

\[ \mathcal{D}_l = \{(x, y) \in \mathcal{X} \times \mathcal{Y}\} \]

Unlabeled data

\[ \mathcal{D}_u = \{x \in \mathcal{X}\} \]

Unsupervised learning: the goal is to recover the structure of the data, eg. clusters, manifolds, low dimensional embeddings, density ...
Why Should SSL Make Sense?

Learner
\[ f(x; \theta) \in \mathcal{F} : \mathcal{X} \rightarrow \mathcal{Y} \]

Labeled data
\[ \mathcal{D}_l = \{(x, y) \in \mathcal{X} \times \mathcal{Y}\} \]

Unlabeled data
\[ \mathcal{D}_u = \{x \in \mathcal{X}\} \]

Unsupervised learning: the goal is to recover the structure of the data, e.g. clusters, manifolds, low dimensional embeddings, density ...

Density
\[ p(x) \]
When unlabeled data is going to help?

Bayes rule:

\[
p(y|x) = \frac{p(y) \cdot p(x|y)}{p(x)}
\]

- **posterior**
- **prior**
- **likelihood**
- **evidence**
When unlabeled data is going to help?

Bayes rule:

\[
p(y|x) = \frac{p(y) \cdot p(x|y)}{p(x)}
\]

posterior

prior

likelihood

evidence

Decision rule:

\[
\hat{y} = \arg\max_k p(k|x) = \arg\max_k \frac{p(k) \cdot p(x|k)}{p(x)} = \arg\max_k p(k \cdot p(x|k))
\]
When unlabeled data is going to help?

- When we expect that $p(x)$ (structure of the data) is related to the $p(y|x)$, i.e. they share parameters.
When unlabeled data is going to help?

- When we expect that $p(x)$ (structure of the data) is related to the $p(y|x)$, i.e., they share parameters.
- In other words, a better estimation of $p(x)$ can improve the estimation of $p(y|x)$. 
When unlabeled data is going to help?

▶ When we expect that $p(x)$ (structure of the data) is related to the $p(y|x)$, i.e., they share parameters.
▶ In other words, a better estimation of $p(x)$ can improve the estimation of $p(y|x)$.
When unlabeled data is going to help?

- When we expect that \( p(x) \) (structure of the data) is related to the \( p(y|x) \), i.e., they share parameters.
- In other words, a better estimation of \( p(x) \) can improve the estimation of \( p(y|x) \).
When unlabeled data is going to help?

- When we expect that $p(x)$ (structure of the data) is related to the $p(y|x)$, i.e. they share parameters.
- In other words, a better estimation of $p(x)$ can improve the estimation of $p(y|x)$. 

![Graph showing the relationship between $p(x)$ and $p(y|x)$](image-url)
Semi-supervised learning often is effective, if the assumptions regarding the relationship between the structure of the data $p(x)$ and the posterior $p(y|x)$ are true for a given problem.
Computer Vision Problems: Object Recognition

Structure: similar images may contain similar objects.
Computer Vision Problems: Object Detection

Structure: similar image patches may contain similar objects. Very close patches (over the 2D image neighborhood) may contain the same object.
Computer Vision Problems: Object Tracking

Structure: similar image patches may contain similar objects. Very close patches (over the 2D image neighborhood and time) may contain the same object.
Structure: similar **pixels** may correspond to the same object. Very close pixels (over the 2D image neighborhood and time) may belong to the same object.
Different Semi-Supervised Learning Settings

- Semi-supervised classification: $f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{Y} = \{1, \cdots, K\}$. 
Different Semi-Supervised Learning Settings

- **Semi-supervised classification**: $f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{Y} = \{1, \ldots, K\}$.
- **Semi-supervised regression**: $f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{Y} = \mathbb{R}$. 
Different Semi-Supervised Learning Settings

- Semi-supervised classification: \( f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{Y} = \{1, \cdots, K\} \).
- Semi-supervised regression: \( f : \mathcal{X} \rightarrow \mathcal{Y}, \mathcal{Y} = \mathbb{R} \).
- Semi-supervised clustering: constrained clustering, clustering with pair-wise must-link and cannot-link constraints.
Transductive and Inductive SSL

- **Transductive:** find $f : \mathcal{D}_l \cup \mathcal{D}_u \rightarrow \mathcal{Y}^{\mathcal{D}_l \cup \mathcal{D}_u}$. Can not be used for any future example which was not in the training set.
- **Inductive:** find $f : \mathcal{X} \rightarrow \mathcal{Y}$. Can be used for any future example, beyond the training set.
Interactive Segmentation
Other Uses of Unlabeled Data

- Unsupervised preprocessing: normalization, standardization, PCA, ICA, ...

[Images from Torralba ICCV 2009, Mobahi et al. ICML 2009, Ranzato NIPS 2007]
Other Uses of Unlabeled Data

- Unsupervised preprocessing: normalization, standardization, PCA, ICA, ...
- Feature extraction: bag-of-words

[Images from Torralba ICCV 2009, Mobahi et al. ICML 2009, Ranzato NIPS 2007]
Other Uses of Unlabeled Data

- Unsupervised preprocessing: normalization, standardization, PCA, ICA, ...

- Feature extraction: bag-of-words

- Unsupervised feature learning: deep learning and sparse coding

[Images from Torralba ICCV 2009, Mobahi et al. ICML 2009, Ranzato NIPS 2007]
The Simplest Approach to SSL

Self-training is a **meta learning (wrapper)** semi-supervised method.

### Self-Training

- **Inputs:** learning algorithm $T$, labeled set $\mathcal{D}_l$, and unlabeled set $\mathcal{D}_u$.
- **For** $n = 1$ to $N$:
  1. Train using the labeled set: $f_n = T(\mathcal{D}_l)$.
  2. Use $f_n$ to classify the unlabeled set: $c_u = f_n(\mathcal{D}_u)$.
  3. Create the set of $m$ most confident examples from the unlabeled set: $C \subseteq \mathcal{D}_u$.
  4. Update the labeled set: $\mathcal{D}_l \leftarrow \mathcal{D}_l \cup \{(x, f_n(x)) | x \in C\}$.
  5. Update the unlabeled set: $\mathcal{D}_u \leftarrow \mathcal{D}_u \setminus C$. 

---

Amir Saffari | Learning with Ambiguities
Example: 1-NN Classifier

(a) Iteration 1

(b) Iteration 25

[Images from Zhu TPCL 2009]
Example: 1-NN Classifier

[Images from Zhu TPCL 2009]
Example: Interactive Segmentation with RF
Example: Self-Training with RF (Iteration 5)
Example: Self-Training with RF (Iteration 10)
Generative Models

Joint Distribution

\[ p(x, y | \theta, \pi) = \underbrace{p(x | y; \theta)}_{\text{mixture component}} \underbrace{p(y | \pi)}_{\pi_y} \]
### Generative Models

**Joint Distribution**

\[
p(x, y | \theta, \pi) = \frac{p(x | y; \theta) p(y | \pi)}{\pi_y} \text{ mixture component}
\]

**Posterior**

\[
p(y | x; \theta, \pi) = \frac{\pi_y p(x | y; \theta)}{\sum_k \pi_k p(x | k; \theta)}
\]
Expectation Maximization Approach

Log-likelihood

\[ \ell(\theta, \pi; D_l, D_u) = \sum_{(x,y) \in D_l} \log \pi_y p(x|y; \theta) + \lambda \sum_{x \in D_u} \log \sum_k \pi_k p(x|k; \theta) \]

- **labeled data**
- **unlabeled data**
Expectation Maximization Approach

Log-likelihood

\[
\ell(\theta, \pi; D_l, D_u) = \sum_{(x,y)\in D_l} \log \pi_y p(x|y; \theta) + \lambda \sum_{x\in D_u} \log \sum_k \pi_k p(x|k; \theta)
\]

Expectation Maximization (EM) can be used to estimate the model parameters.

**EM**

- **Inputs:** labeled set \(D_l\), and unlabeled set \(D_u\), initial model parameters \(\theta_0, \pi_0\).
- **For** \(n = 1\) to \(N\):
  1. **E step:** Estimate the posterior \(p(y|x; \theta_{n-1}, \pi_{n-1})\) for \(D_u\).
  2. **M step:** Update the model parameters, given the posteriors for unlabeled data: \(\theta_n, \pi_n = \arg\max_{\theta, \pi} \ell(\theta, \pi; D_l, D_u)\).
Example: Two Gaussians
Example: Supervised GMM
Example: Semi-Supervised EM with GMM
Assumptions

[Images from Jaakkola et al.]
Assumptions

[Images from Jaakkola et al.]
Maximum Margin Linear Classifier

\[ wx - b = 0 \]

\[ wx - b = 1 \]
S3VM and TSVM
S3VM

SVM with soft margins: binary case $y \in \{-1, 1\}$ without the bias term

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y \langle w, x \rangle)$$

- margin
- hinge loss
SVM with soft margins: binary case $y \in \{-1, 1\}$ without the bias term

$$\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y \langle w, x \rangle)$$

Prediction rule:

$$\hat{y} = \text{sign}(\langle w, x \rangle)$$

Margin of an unlabeled data:

$$m_u(w, x) = |\langle w, x \rangle|$$
S3VM: Loss Functions
S3VM: Formulation

S3VM

\[
\min_{\mathbf{w}} \frac{\lambda}{2} \left\| \mathbf{w} \right\|_2^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y \langle \mathbf{w}, \mathbf{x} \rangle) + \frac{\gamma}{|D_u|} \sum_{\mathbf{x} \in D_u} \max(0, 1 - |\langle \mathbf{w}, \mathbf{x} \rangle|)
\]

- Margin
- Hinge loss
- Sym. hinge
S3VM: Formulation

\[
\begin{align*}
\text{S3VM} & \quad \min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y \langle w, x \rangle) + \\
& \quad \quad \left( \text{margin} \right) \quad \text{hinge loss} \\
& \quad \quad + \frac{\gamma}{|D_u|} \sum_{x \in D_u} \max(0, 1 - |\langle w, x \rangle|) \quad \left( \text{sym. hinge} \right) \\
\text{Balancing constraint:} & \quad \frac{1}{|D_l|} \sum_{(x,y) \in D_l} y = \frac{1}{|D_u|} \sum_{x \in D_u} \langle w, x \rangle
\end{align*}
\]

[Vapnik 1998, Joachims ICML 1999]
Entropy Regularization

Logistic Regression

\[ p(y|x; w) = \frac{1}{1 + e^{-y\langle w, x \rangle}} \]
Entropy Regularization

Logistic Regression

\[ p(y|x; w) = \frac{1}{1 + e^{-y\langle w, x \rangle}} \]

Log-likelihood:

\[ \ell(w; D_1) = \sum_{(x,y) \in D_1} \log p(y|x; w) - \frac{\lambda}{2} \| w \|_2^2 \]
Entropy Regularization

Logistic Regression

\[ p(y|x; w) = \frac{1}{1 + e^{-y\langle w, x \rangle}} \]

Log-likelihood:

\[ \ell(w; D_1) = \sum_{(x,y) \in D_1} \log p(y|x; w) - \frac{\lambda}{2} \| w \|_2^2 \]

Entropy Minimization: (semi-supervised logistic regression)

\[ \min_w \frac{\lambda}{2} \| w \|_2^2 + \sum_{(x,y) \in D_1} \log(1 + e^{-y\langle w, x \rangle}) - \gamma \sum_{x \in D_u} \sum_k p(k|x; w) \log p(k|x; w) \]

[Grandvalet and Bengio, 2006]
Entropy Regularization

- Hinge
- Sym. Hinge
- Logit
- Entropy
Example: Interactive Segmentation with Linear SVM
Example: Interactive Segmentation with Linear TSVM
Example: Linear TSVM

S3VM in local minimum

S3VM in wrong gap

[Images from Zhu TPCL 2009]
Cluster Assumption and Margin Maximization
Cluster Kernels

Linear SVM:

\[ f(x) = \langle w, x \rangle \]
Cluster Kernels

**Linear SVM:**

\[ f(x) = \langle w, x \rangle \]

**Nonlinear SVM:**

\[ f(x') = \langle w, \phi(x') \rangle = \sum_{(x,y) \in \mathcal{D}_l} y \alpha_x \langle \phi(x), \phi(x') \rangle K(x,x') \]

nonlinear mapping
Cluster Kernels

Linear SVM:

\[ f(x) = \langle w, x \rangle \]

Nonlinear SVM:

\[ f(x') = \langle w, \phi(x') \rangle = \sum_{(x,y) \in D_l} y\alpha_x \langle \phi(x), \phi(x') \rangle K(x,x') \]

Cluster Kernel:

\[ K_c(x, x') = K(x, x') \frac{\sum_p \mathbb{I}(c(x) == c(x'))}{n} \]

[Weston et al. NIPS 2003]
Boosting

$$f(x) = \sum_{m=1}^{M} w_m g(x; \theta_m)$$
Boosting

$$f(x) = \sum_{m=1}^{M} w_m g(x; \theta_m)$$

Base Learner

$$g(x; \theta_m) \in G : \mathcal{X} \rightarrow \mathcal{Y}$$
Boosting

\[
f(x) = \sum_{m=1}^{M} w_m \, g(x; \theta_m)
\]

Base Learner

\[
g(x; \theta_m) \in G : \mathcal{X} \rightarrow \mathcal{Y}
\]

Boosting is a linear classifier over the space of base learners \(G\):

\[
f(x) = G(x; \theta)w
\]
Multi-Class Boosting and Margin

\[ g(x; \theta_2) \]

\[ w \]

\[ G_{y,..}(x, \theta) \]

\[ m(x, y; f) \]

\[ G_{y',..}(x, \theta) \]

\[ g(x; \theta_1) \]
Boosting: Learning with Functional Gradient Descent

\[ f(x; \beta^*) = \arg \min_{\beta} \frac{1}{|\mathcal{X}_l|} \sum_{(x, y) \in \mathcal{X}_l} \ell(x, y; f) \Rightarrow f(x; \beta^*) = \sum_{m=1}^{M} w^*_m \ g(x; \theta^*_m) \]
Boosting: Learning with Functional Gradient Descent

\[
f(x; \beta^*) = \arg \min_{\beta} \frac{1}{|\mathcal{X}_l|} \sum_{(x, y) \in \mathcal{X}_l} \ell(x, y; f) \Rightarrow f(x; \beta^*) = \sum_{m=1}^{M} w_m^* g(x; \theta_m^*)
\]
Boosting with Prior

Priors

∀x ∈ Du, k ∈ Y : q(k|x)

\[
\min_{w,\theta} \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \ell(x, y; w, \theta) + \supervised
\]

\[
+ \frac{\gamma}{|D_u|} \sum_{x \in D_u} j^p(x, q; w, \theta) \text{ Prior}
\]

Robust Loss Functions

[Saffari et al. ECCV 2010]
Cluster Prior

\[
p(y|x) = \text{[red, blue, green]} 
\]

[Saffari et al. CVPR 2009]
Cluster Prior

$p(y|x) = \text{[red, blue, green]}$

[Saffari et al. CVPR 2009]
Cluster Prior

[Saffari et al. CVPR 2009]

Amir Saffari

Learning with Ambiguities
Example: Interactive Segmentation with Boosting

original

label

prior

segmentation
Example: Interactive Segmentation with Cluster Prior
Manifold Assumption

[Images from Hein and von Luxburg MLSS 2007]
Manifold Assumption

[Images from Hein and von Luxburg MLSS 2007]
Manifold Assumption

[Images from Hein and von Luxburg MLSS 2007]
## Label Propagation: Supervised

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asteroid</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bright</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comet</td>
<td></td>
<td></td>
<td>●</td>
<td></td>
</tr>
<tr>
<td>year</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>zodiac</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>airport</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>bike</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>camp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yellowstone</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>zion</td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

[Images from Zhu TPCL 2009]
Label Propagation: Semi-Supervised

<table>
<thead>
<tr>
<th></th>
<th>$d_1$</th>
<th>$d_5$</th>
<th>$d_6$</th>
<th>$d_7$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_8$</th>
<th>$d_9$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>asteroid</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bright</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comet</td>
<td></td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>year</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zodiac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>airport</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bike</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>camp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yellowstone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
<tr>
<td>zion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>●</td>
</tr>
</tbody>
</table>

[Images from Zhu TPCL 2009]
Label Propagation: Semi-Supervised

[Images from Kveton et al. CVPR OLCV 2010]
Label Propagation: Semi-Supervised

[Images from Kveton et al. CVPR OLCV 2010]
Mincut

\[
\begin{align*}
\text{Mincut energy:} & \quad \min_{\{y \mid x \}} \sum_{x \in D} u_s(x, x') (y_x - y_{x'})^2 + \sum_{x' \in D} u_s(x, x') (y_x - y_{x'})^2
\end{align*}
\]
Mincut energy:

\[
\min_{\{y_x\}} \sum_{x \in D_u} \left( \sum_{(x', y') \in D_l} s(x, x')(y_x - y')^2 + \sum_{x' \in D_u} s(x, x')(y_x - y_{x'})^2 \right)
\]
Random Walks on Graph

[Images from Zhu TPCL 2009]
LapSVM

\[
\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y \langle w, x \rangle) + \\
+ \frac{\gamma}{|D_u|^2} \sum_{x \in D_u} \sum_{x' \in D_l \cup D_u} s(x, x')(\langle w, x \rangle - \langle w, x' \rangle)^2
\]
Manifold Regularization: LapSVM

LapSVM

\[
\min_{w} \frac{\lambda}{2} \|w\|^2 + \frac{1}{|\mathcal{D}_l|} \sum_{(x, y) \in \mathcal{D}_l} \max(0, 1 - y \langle w, x \rangle) +
\]

\[
+ \frac{\gamma}{|\mathcal{D}_u|^2} \sum_{x \in \mathcal{D}_u} \sum_{x' \in \mathcal{D}_l \cup \mathcal{D}_u} s(x, x')(\langle w, x \rangle - \langle w, x' \rangle)^2
\]

Graph Laplacian

\[
\min_{w} \frac{\lambda}{2} \|w\|^2 + \frac{1}{|\mathcal{D}_l|} \sum_{(x, y) \in \mathcal{D}_l} \max(0, 1 - y \langle w, x \rangle) +
\]

\[
+ \frac{\gamma}{|\mathcal{D}_u|^2} \langle f, Lf \rangle
\]

where \( f \) is the response vector of the classifier to both labeled and unlabeled data.

[Belkin et al. JMLR 2006]
Boosting with Priors and Manifolds

\[
\min_{w, \theta} \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \ell(x, y; w, \theta) + \text{Supervised} \\
+ \frac{\gamma}{|D_u|} \sum_{x \in D_u} \left( \lambda j^p(x, q; w, \theta) + (1 - \lambda) \sum_{x' \in D_u, x' \neq x} \frac{s(x, x')}{z(x)} j^m(x, x'; w, \theta) \right) \text{Prior} \\
+ \text{Unsupervised}
\]
Example: Boosting with Cluster and Manifold Priors

<table>
<thead>
<tr>
<th>Methods-Dataset</th>
<th>g241c</th>
<th>g241d</th>
<th>Digit1</th>
<th>USPS</th>
<th>COIL</th>
<th>BCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>44.05</td>
<td>43.22</td>
<td>23.47</td>
<td>19.82</td>
<td>65.91</td>
<td>48.74</td>
</tr>
<tr>
<td>SVM</td>
<td>47.32</td>
<td>46.66</td>
<td>30.60</td>
<td>20.03</td>
<td>68.36</td>
<td>49.85</td>
</tr>
<tr>
<td>RF (weak)</td>
<td>47.51</td>
<td>48.44</td>
<td>42.42</td>
<td>22.90</td>
<td>75.72</td>
<td>49.35</td>
</tr>
<tr>
<td>GBoost-RF</td>
<td>46.77</td>
<td>46.61</td>
<td>38.70</td>
<td>20.89</td>
<td>69.85</td>
<td>49.12</td>
</tr>
<tr>
<td>T SVM</td>
<td>24.71</td>
<td>50.08</td>
<td>17.77</td>
<td>25.20</td>
<td>67.50</td>
<td>49.15</td>
</tr>
<tr>
<td>Cluster Kernel</td>
<td>48.28</td>
<td>42.05</td>
<td>18.73</td>
<td>19.41</td>
<td>67.32</td>
<td>48.31</td>
</tr>
<tr>
<td>LapSVM</td>
<td>46.21</td>
<td>45.15</td>
<td>08.97</td>
<td>19.05</td>
<td>N/A</td>
<td>49.25</td>
</tr>
<tr>
<td>ManifoldBoost</td>
<td>42.17</td>
<td>42.80</td>
<td>19.42</td>
<td>19.97</td>
<td>N/A</td>
<td>47.12</td>
</tr>
<tr>
<td>SemiBoost-RF</td>
<td>48.41</td>
<td>47.19</td>
<td>10.57</td>
<td>15.83</td>
<td>63.39</td>
<td>49.77</td>
</tr>
<tr>
<td>MCSSB-RF</td>
<td>49.77</td>
<td>48.57</td>
<td>38.50</td>
<td>22.95</td>
<td>69.96</td>
<td>49.12</td>
</tr>
<tr>
<td>GPMBoost-RF</td>
<td>15.69</td>
<td>39.45</td>
<td>11.19</td>
<td>14.92</td>
<td>62.60</td>
<td>49.27</td>
</tr>
</tbody>
</table>

[Saffari et al. 2010]
## Example: Boosting with Cluster and Manifold Priors

<table>
<thead>
<tr>
<th>Methods-Dataset</th>
<th>g241c</th>
<th>g241d</th>
<th>Digit1</th>
<th>USPS</th>
<th>COIL</th>
<th>BCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-NN</td>
<td>40.28</td>
<td>37.49</td>
<td>06.12</td>
<td>07.64</td>
<td>23.27</td>
<td>44.83</td>
</tr>
<tr>
<td>SVM</td>
<td>23.11</td>
<td>24.64</td>
<td>05.53</td>
<td>09.75</td>
<td>22.93</td>
<td>34.31</td>
</tr>
<tr>
<td>RF (weak)</td>
<td>44.23</td>
<td>45.02</td>
<td>17.80</td>
<td>16.73</td>
<td>34.26</td>
<td>43.78</td>
</tr>
<tr>
<td>GBoost-RF</td>
<td>31.84</td>
<td>32.38</td>
<td>06.24</td>
<td>13.81</td>
<td>21.88</td>
<td>40.08</td>
</tr>
<tr>
<td>TSVM</td>
<td>18.46</td>
<td>22.42</td>
<td>06.15</td>
<td>09.77</td>
<td>25.80</td>
<td>33.25</td>
</tr>
<tr>
<td>Cluster Kernel</td>
<td>13.49</td>
<td>04.95</td>
<td>03.79</td>
<td>09.68</td>
<td>21.99</td>
<td>35.17</td>
</tr>
<tr>
<td>LapSVM</td>
<td>23.82</td>
<td>26.36</td>
<td>03.13</td>
<td>04.70</td>
<td>N/A</td>
<td>32.39</td>
</tr>
<tr>
<td>ManifoldBoost</td>
<td>22.87</td>
<td>25.00</td>
<td>04.29</td>
<td>06.65</td>
<td>N/A</td>
<td>32.17</td>
</tr>
<tr>
<td>SemiBoost-RF</td>
<td>41.26</td>
<td>39.14</td>
<td>02.56</td>
<td>05.92</td>
<td>15.31</td>
<td>47.12</td>
</tr>
<tr>
<td>MCSSB-RF</td>
<td>45.11</td>
<td>40.26</td>
<td>13.19</td>
<td>12.31</td>
<td>31.09</td>
<td>47.64</td>
</tr>
<tr>
<td>GPMBoost-RF</td>
<td>12.80</td>
<td>12.59</td>
<td>02.35</td>
<td>06.33</td>
<td>14.49</td>
<td>45.41</td>
</tr>
</tbody>
</table>

[Saffari et al. 2010]
Example: Interactive Segmentation with Cluster Prior
Example: Interactive Segmentation with Cluster Prior and Manifolds
Learning from Different Views

- Data is represented by multiple views: \( x = [x_1^T | \cdots | x_V^T]^T \)
Learning from Different Views

- Data is represented by multiple views: \( x = [x_1^T | \cdots | x_V^T]^T \)
- There is a classifier per view: \( \mathcal{F} = \{f_v\}_{v=1}^V \)

Multi-View Learning

\[
\mathcal{F}^* = \arg \min_{\mathcal{F}} \sum_{(x,y) \in \mathcal{D}_1} \ell(x, y; \mathcal{F}) + \gamma \sum_{x \in \mathcal{D}_u} \phi(x; \mathcal{F})
\]
Co-Training

- Inputs: learning algorithm $T$, labeled set $D_l$, and unlabeled set $D_u$.
- For $n = 1$ to $N$:
  1. Train using the labeled set: $f^1_n = T(D^1_l)$ and $f^2_n = T(D^2_l)$.
  2. Use $f^1_n$ and $f^2_n$ to classify the unlabeled set: $c^1_u = f^1_n(D^1_u)$ and $c^2_u = f^2_n(D^2_u)$.
  3. Create the set of $m$ most confident examples from each view of the unlabeled set: $C^1 \subset D^1_u$ and $C^2 \subset D^2_u$.
  4. Update the labeled set:
     $D_l \leftarrow D_l \cup \{(x, f^1_n(x)) | x \in C^1\} \cup \{(x, f^2_n(x)) | x \in C^2\}$.
  5. Update the unlabeled set: $D_u \leftarrow D_u \setminus (C^1 \cup C^2)$.

[Blum and Mitchell COLT 1998]
Co-Training and Assumptions

- We can represent the data in two views: \( x = [x^1, x^2] \).
Co-Training and Assumptions

- We can represent the data in two views: $x = [x^1, x^2]$.
- We can train a good classifier only using either $x^1$ or $x^2$.

[Blum and Mitchell COLT 1998, Sindhwani et al. ICML 2005]
Co-Training and Assumptions

- We can represent the data in two views: \( x = [x^1, x^2] \).
- We can train a good classifier only using either \( x^1 \) or \( x^2 \).
- \( x^1 \) and \( x^2 \) are conditionally independent given the class.

[Blum and Mitchell COLT 1998, Sindhwani et al. ICML 2005]
Co-Training and Assumptions

- We can represent the data in two views: \( x = [x^1, x^2] \).
- We can train a good classifier only using either \( x^1 \) or \( x^2 \).
- \( x^1 \) and \( x^2 \) are conditionally independent given the class.

[Blum and Mitchell COLT 1998, Sindhwani et al. ICML 2005]
Multi-View Boosting with Priors

Multi-View Priors

\[ q_v(k|x_v) = \frac{1}{V-1} \sum_{s \neq v} p_s(k|x_s), \quad \forall v \in \{1, \cdots, V\}, k \in \mathcal{Y} \]
Multi-View Boosting with Priors

Multi-View Priors

\[
q_v(k|x_v) = \frac{1}{V-1} \sum_{s \neq v} p_s(k|x_s), \quad \forall v \in \{1, \cdots, V\}, k \in \mathcal{Y}
\]

Boosting with Priors:

\[
\arg\min_{w, \theta} \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \ell(x, y; w, \theta) + \frac{\gamma}{|D_u|} \sum_{x \in D_u} j^p(x, q; w, \theta)
\]

[Refs: Saffari et al. ECCV 2010]
Example: Interactive Segmentation with RF

original

label

prior

segmentation
Example: Interactive Segmentation with Linear SVM
Example: Interactive Segmentation with Multi-View Classifiers
Multiple Instance Learning
miSVM

\[
\min_{\{y_{ij}\}} \min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y f_w(x)) \\
\text{s.t. } \forall j \in B_i^+: y_{ij} \in \{-1, +1\}, \sum_j \frac{y_{ij} + 1}{2} \geq 1.
\]

[Andrews et al. NIPS 2003]
From SSL to MIL

\[
\min_w \frac{\lambda}{2} \|w\|^2 + \frac{1}{|D_l|} \sum_{(x,y) \in D_l} \max(0, 1 - y\langle w, x \rangle) + \\
\max \left(0, 1 - \frac{\sum_{x \in D_u} \max(0, 1 - |\langle w, x \rangle|)}{|D_u|} \right)
\]

\text{s.t. } \forall j \in B_i^+ : \max_j \langle w, x_{ij} \rangle \geq 0.

MI-SVM (Learning with Latent Variables)

MI-SVM (Latent SVM)

\[ f_w(x) = \max_z \langle w, \phi(x, z) \rangle \]

\[
\min_w \frac{\lambda}{2} \| w \|^2_2 + \frac{1}{|D_l|} \sum_{(x, y) \in D_l} \max(0, 1 - yf_w(x))
\]

[Felzenszwalb et al. CVPR 2008]
Transfer Learning

(a) Traditional Machine Learning

(b) Transfer Learning

[Images from Pan and Yang TKDE 2009]
Transfer Learning

- Inductive Transfer Learning
  - Labeled data are available in a target domain
  - Labeled data are available in a source domain
  - No labeled data in a source domain

- Transductive Transfer Learning
  - Labeled data are available only in a source domain
  - Assumption: different domains but single task

- Unsupervised Transfer Learning
  - No labeled data in both source and target domain
  - Assumption: single domain and single task

- Multi-task Learning
  - Source and target tasks are learned simultaneously

- Domain Adaptation
  - Assumption: single domain and single task

Case 1
- No labeled data in a source domain
- Self-taught Learning

Case 2
- Assumption: different domains but single task
- Sample Selection Bias / Covariance Shift

[Images from Pan and Yang TKDE 2009]
Active Learning is a form of supervised machine learning in which the learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new unlabeled data points.

▶ Self-training vs uncertainty sampling
Active Learning

Active learning is a form of supervised machine learning in which the learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new unlabeled data points.

- Self-training vs uncertainty sampling
- Co-training vs query-by-committee
Active learning is a form of supervised machine learning in which the learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at new unlabeled data points.

- Self-training vs uncertainty sampling
- Co-training vs query-by-committee
- Exploitation vs Exploration